homework 3 report

To evaluate my model, I used a 4-fold cross validation method. I felt that because the holdout method relies on an arbitrary partition of the data, the evaluation would depend too heavily on which points ended up in the training set and which ended up in the test set, so evaluation may have been significantly different depending on how the division was made. I also did not want to use bootstrap because I wanted all the training tuples to be unique when I used them to build the decision tree. Therefore, I decided on doing 4 rounds of cross-validation. I used a smaller number of folds for the sake of efficiency.

Here are the metrics I got using my model. Overall, they were all somewhat similar in value, hovering around the 70-75% range, implying that my algorithm doesn't seem to have a large "bias" in the number of false positives it identifies versus false negatives, and vice versa. However, because the specificity value was smallest, it's possible that my model has a tendency to identify false positives. In general, I would prefer my metrics to be higher, since predicting heart disease is an important task with large consequences in the real world. Thus, this assignment helped me understand the importance of building good classification models, as well as the importance of testing these models, because evaluating the quality of my induction tree really helped me understand just how effective (or not effective) it was, and if I were to continue improving it in the future, these numbers would be very useful in helping guide the adjustments I could make to the algorithm.

- accuracy:

$$\frac{1}{4}\left(\frac{59}{75}+\frac{45}{75}+\frac{56}{75}+\frac{56}{78}\right)=.713$$

  This indicates that on average, 71.3% of the test tuples were identified correctly. This percentage isn't terrible, but it also isn't as high as I'd like it to be; ideally, more test tuples would be correctly identifies by my model.

- precision:

$$\frac{1}{4}\left(\frac{32}{39}+\frac{28}{44}+\frac{32}{46}+\frac{27}{32}\right)=.749$$

  This means 74.9% of tuples labeled positive were actually positive. There were more false positives than I would like, especially because in the real world if someone was misclassified as having heart disease, it could cause a lot of stress and inconvenience even if the patient eventually realizes they were misdiagnosed.

- recall or sensitivity:

$$\frac{1}{4}\left(\frac{32}{41}+\frac{28}{42}+\frac{32}{37}+\frac{27}{44}\right)=.731$$

  73.1% of positive tuples were labeled as positive. Because recall is lower than sensitivity, it indicates that there were more false negatives identified on average than false positives. Again, I would like the

recall rate to be much higher than it is, because the consequences of a false negative in this context (i.e. someone with heart disease not being diagnosed) could be devastating if the disease isn't caught and the patient therefore doesn't do anything to treat it.

- F:

$$\frac{1}{4}\left(\frac{2(.8205)(.7805)}{.8205+.7805}+\frac{2(2/3)(.63636)}{2/3+.63636}+\frac{2(.69565)(.86486)}{.69565+.86486}+\frac{2(.84375)(.6136)}{.84375+.6136}\right)=.733$$

This gives the harmonic mean of precision and recall, and is in between the values of the precision and recall given above, though leaning more toward the recall value.

- specificity:

$$\frac{1}{4}\left(\frac{27}{34}+\frac{17}{33}+\frac{24}{38}+\frac{29}{34}\right)=.698$$

69.8% of negative tuples were correctly identified as negative. This is lower than the sensitivity (recall) value above, indicating that positive tuples were more likely to be correctly identified than negative tuples.